A Computational Normative Theory of Scientific Evidence

David B. Sher
Computer Science Department, State University of New York at Buffalo, New York

ABSTRACT

A scientific reasoning system makes decisions using objective evidence in the form of independent experimental trials, propositional axioms, and constraints on the probabilities of events. I propose a collection of algorithms that derive probability intervals and estimate conditional probabilities from objective evidence in those forms. This reasoning system can manage uncertainty about data and rules in a rule-based expert system. I expect that the system will be particularly applicable to diagnosis and analysis in domains with a wealth of experimental evidence such as medicine. The algorithms currently apply to systems with arbitrary amounts of experimental evidence but with less than 20 variables. I discuss limitations of this solution and propose future directions for this research. This work can be considered a generalization of Nilsson's "probabilistic logic" to intervals and experimental observations.

KEYWORDS: ***interval probability, evidence combination, experimental evidence, probabilistic logic, statistical inference***

1. INTRODUCTION

Expert systems were originally defined as a method of encoding the knowledge and reasoning of a human expert into a computer program. However, for many diagnostic and analytical reasoning problems we would prefer that the computer program base its reasoning on objective facts rather than human expertise, especially when no experts are available or the reasoning of the available experts is suspect. Thus, I propose a system for reasoning from objective criteria—experiments performed by the method of identical

Address correspondence to David Sher, Computer Science Department, SUNY Buffalo, Buffalo, NY 14260.

Received July 1, 1990; accepted September 15, 1991.

International Journal of Approximate Reasoning 1992; 6:505–524
© 1992 Elsevier Science Publishing Co., Inc.
655 Avenue of the Americas, New York, NY 10010 0888-613X/92/$5.00

independent trials. Since identical independent trials are a standard method of gathering scientific data and there is a vast store of literature containing the results of this type of experiments, I feel that a system that reasons accurately from such data will be of great utility.

When objective data are collected by scientists and engineers they are often in the form of independently distributed trials or experiments. An example of experimental data is, In a sample of 100 toys, 10 arrived damaged. Some of these experiments verify facts, as in this example; other experiments verify rules such as 10% of 20 red toys arrived damaged; this experiment establishes a relationship between redness and damage. Several rules can conflict, such as 90% of 10 toy trucks arrived damaged, which raises the question of the red toy truck.

The system presented here does not attempt to simulate the reasoning of an expert or a scientist. Rather, it is a normative system that attempts to generate, or at least bound, the objectively correct values for the requested probabilities.

This paper describes a computationally feasible algorithm for making decisions from this type of evidence in small or simple domains (domains that can be represented using less than 20 variables). I present a preliminary methodology for evidence combination, where the source of the evidence is experimental in nature.

I assume throughout the paper that the frequencies provided to the system were collected using correct scientific methodology and thus they represent insofar as possible independent and independently distributed scientific trials. Representing and reasoning from scientific data collected using a flawed methodology would require a far more complex theory, which should be the subject of future work.

The scheme presented here can also account for other methods of expressing objective information such as axiomatic statements such as Trucks have wheels, and restrictions on the probabilities of events, such as The probability of heads on a fair coin is 0.5, the probability of heads when Joe is flipping the coin is somewhere between 0.6 and 0.8.

The strength of this system is its ability to accept evidence in a large variety of formats that can express many different forms of uncertainty. Initially I will focus on our management of experimental evidence, since that is the unique feature of our approach.

2. PREVIOUS WORK

This work was inspired largely by Henry Kyburg and Ron Loui's work [1-4] at the University of Rochester on the logical foundations of statistical inference. Their work uses experimental results to derive sets of confidence intervals and then develops an algebra of these sets of confidence intervals.

Kyburg [1] discussed the use of second-order probability distributions for subjective evidence evaluations. Such an approach requires subjective evaluations of an exponential number of probabilities. He did not investigate objective higher order probabilities, but he did suggest objective interval probabilities as most appropriate for objective uncertainty management.

Another important influence on this work is the Dempster–Shafer system [5, 6] for managing uncertainty and the analyses of it [7–9]. Dempster–Shafer reasoning expresses belief as mass functions over sets of possible beliefs and develops a calculus of such mass functions for evidence combination. Although such mass functions may be a good way of summarizing human expertise, I felt that they were not suitable for expressing the uncertainty inherent in objective evidence. Hau and Kashyap [10] suggested a method of applying Dempster–Shafer and fuzzy reasoning to rule-based expert systems that consider one of the best approaches to this problem if the rules and facts in the system are derived from a subjective or qualitative source.

This work owes a great deal to the Bayesian interval approach to uncertainty; it reduces to Good's Bayesian interval approach [11, 12] when only interval probabilities are supplied. Our work also is meant to handle the same class of applications as Judea Pearl's probability networks [13] and other Bayesian network paradigms [14–24]. Of particular note on this topic is Lauritzen and Spiegelhalter's work [22] on efficient calculation of probabilities from Bayesian belief networks. Dr. Cheeseman notes [25] that difficulties are presented by the number of probabilities that must be entered into such a system. He is skeptical that all of these numbers can be accurately estimated. As our system uses observations rather than probabilities as its input, we avoid these inaccuracies.

Herskovits and Cooper [23] study the extraction of a belief network from a large database of clinical cases. However, this paper assumes that the database already contains accurate conditional probabilities for event and thus suffers from the difficulties noted by Cheeseman. The presence of accurate conditional probabilities does not seem to be necessary to this approach, and perhaps the same approach can be applied to accurately estimating a network from observations or trials.

Our work generalizes Nilsson's "probabilistic logic" [26], where he also studies mapping joint distributions into a vector space. His work analyzes the case where all the information items are point probabilities for events; our work reduces to his when our information is in this form.

Our work applies to the same domain as that discussed by Birnbaum [27, 28]. His work, like ours, focuses on likelihoods, as he discusses independent, identically distributed binary experiments, but his work does not consider computational concerns or multiple types of experiments. Our work is different from this and other statistical approaches in that most statistical work studies how to construct an experiment to accurately determine the probability of an

event, whereas our work focuses on what can be deduced from experimental evidence a posteriori. This work is important because not every decision maker has the opportunity to construct experiments relevant to his or her decision; a doctor makes decisions based on published experimental research and the symptoms at hand, but he does not perform new controlled studies for each patient.

Berger has discussed extensively the sufficiency of the likelihood function for characterizing evidence in a monograph with Wolpert [29] and more concisely in his textbook [30]. He presents and justifies as a basis for statistical inference the *likelihood principle*—"All the information about θ obtainable from an experiment is contained in the likelihood function for the actual observation." However, when discussing implementations of the likelihood principle he strongly supports a Bayesian approach. The lack of a truly noninformative prior over joint distributions[1] and computational concerns has led us to a maximum likelihood approach.

Kyburg [3] and Loui [4] have surveyed and discussed a large variety of methods for uncertain inference including most of the systems that are comparable to ours.

3. DEFINITIONS

The input to our system consists of a set of propositions and a set of experiments that test logical combinations of these propositions. For example, consider playing poker against Harry. What can we deduce from Harry's body language about his hand? Consider these two propositions:

A: Harry lit his pipe.

B: Harry has two pairs.

Some experiments about the relationship between these two events are

1. In the first 30 hands Harry lit his pipe in 9 of them.
2. In the next 40 hands Harry had two pairs in 5 of them.
3. In the following hands you noticed that of the 6 times he lit his pipe 5 of those times he had two pairs.

The joint distribution of events in our domain contain all the information about the domain that is useful for deduction. The joint distribution is the probability distribution over the elements of the truth table; that is, in our

[1] Using the uniform prior distribution results in marginal second-order prior distributions that are strongly biased towards probability 1/2. I am unaware of a well-defined prior that has unbiased second-order marginal distributions.

poker example,

- a: A and B
- b: A and Not B
- c: Not A and B
- d: Not A and Not B

If we knew the probability of a, b, c, and d, we could determine the probability of any logical combination of events and all the conditional probabilities too. Hence, to engage in deduction from experimental evidence, we will study the issue of estimating a joint distribution from experimental evidence.

4. MAXIMUM LIKELIHOOD ESTIMATION

In this work I propose estimating the probability of a proposition as one of the probabilities derived from a joint distribution that maximizes the probability of the experimental results.

In estimation theory, if we are estimating the value of e, the function that maps values of e into probabilities of the observed data is called the *likelihood function*. The value of e that maximizes the likelihood function is called the *maximum likelihood estimate*. The maximum likelihood joint distribution is the joint distribution that maximizes the probability of the experimental results.

We assume that each experiment is identically independently distributed (iid) and that the sampling was unbiased. The probability of observing a logical combination of the primitives (such as "A or B") is a sum of probabilities in the joint distribution; in our poker example the probability of A is $a + b$ and that of B is $a + c$. Thus the probability of an experiment consisting of iid trials on A where A was observed 3 out of 5 times is then $(a + b)^3(c + d)^2$. The iid assumption implies that the probability of a set of experiments is the product of the probabilities from each experiment. Hence, the likelihood function for statements 1 and 2 of our poker example is

$$(a + b)^9(c + d)^{21}(a + c)^5(b + d)^{35}$$

Conditional experiments, where the result is reported only when a condition applies, that is statement 3, have a conditional probability. The probability of a conditional experiment is the ratio of the probability of the conjunction of the condition and the event and the probability of the condition. Thus the probability of statement 3 is

$$a^5b/(a + b)^6$$

The probability of statements 1, 2, and 3 is

$$a^5b(a + b)^3(c + d)^{21}(a + c)^5(b + d)^{35}$$

Since we had an experiment on A where A occurred 9 times, we could perform up to nine conditional experiments on A and still express the likelihood function as a polynomial (rather than a rational function). This paper's analysis is restricted to sets of experiments whose likelihood function is a polynomial; this means that given N experiments using condition X, an experiment was done where condition X occurred at least N times. Whether a set of experiments fits this condition can be determined in time $L \log L$, where L is the size of the likelihood function.

5. OBSERVATION SPACE

Each joint distribution is an assignment of probabilities to a finite set of mutually exclusive events; hence a joint distribution can be considered a vector (JDV), and the locus of joint distributions in a set of points in a vector space. In our poker example there are four mutually exclusive events, a, b, c, and d; assignments of probabilities to these events correspond to four-dimensional vectors. For example, if A and B are independent events with probabilities 0.3125 and 0.1429, respectively, then the corresponding JDV would be ($a = 0.0446$, $b = 0.2679$, $c = 0.0982$, $d = 0.5893$).

The probability of an observation is a linear function of the JDV; in our poker example, A has a probability of $a + b$. These linear functions define a dual space of observation vectors (OVs) with the same dimensionality, whose coefficients are 1 if the corresponding element of the joint distribution is compatible with the observation and 0 otherwise. In our poker example, A is represented by the OV (1, 1, 0, 0), and B is (1, 0, 1, 0), and A but not B is the OV (0, 1, 0, 0). The probability of an observation under a joint distribution is the dot product of the OV with the JDV.

Note that the dimensionality of an observation is the same as that of a joint distribution. Hence observation vectors and JDVs can be embedded in the same vector space.

A set of observations resulting from experiments corresponds to a set of OVs that span a vector space called the *observation space*. This space will often be lower dimensional than the space spanned by JDVs. For example, consider the evidence from statements 1 and 2 of the poker example; the observations correspond to these four vectors:

Proposition	Vector
A	(1, 1, 0, 0)
Not A	(0, 0, 1, 1)
B	(1, 0, 1, 0)
Not B	(0, 1, 0, 1)

These vectors span a three-dimensional space. The vector $(1, -1, -1, 1)$ is perpendicular to all four vectors. Hence adding a multiple of $(1, -1, -1, 1)$ to a JDV will not change the probability of these observations. For example, under the JDVs (0.25,0.25, 0.25, 0.25) and (0.5, 0, 0, 0.5), the probabilities of the four observations are the same.

The space spanned by the JDVs is the cross product of the observation space and the space spanned by vectors perpendicular to the observation space. We call the space spanned by vectors perpendicular to the observation space the *null space*. Adding a vector from the null space to a JDV leaves the probability of the observations unchanged.

Note that if the difference of two JDVs for a set of experiments is an element of the null space, then the probability of each observation is the same for both JDVs. Since the probability of all the observations is the product of the probabilities of each observation, the probability of the results from the set of experiments is the same for both JDVs. Hence, the results from those experiments cannot tell us which of the corresponding joint distributions is a better model. For example, no set of unconditional experiments on statements A or B can tell us whether (0.25, 0.25, 0.25, 0,25) or (0.5, 0, 0, 0.5) is a better estimate of the true joint distribution.

6. CONCAVITY AND UNIQUENESS OF THE MAXIMA

Two different joint distributions maximize the likelihood of a set of experiments only if their JDVs differ by an element of the null space for that set of experiments. Hence if we can find a single maximum likelihood joint distribution we know what all of them look like.

First, I show that the likelihood function, L, is concave. Assume that j_1 and j_2 are two JDVs; let $f(\gamma) = \log\{L[j_1 + \gamma(j_2 - j_1)]\}$; $L(x) = \Pi_i s_i(x)$, where s_i are sums of elements of the joint distribution. Each s_i is the dot product of an observation vector and a JDV; hence $L = \Pi_i o_i \cdot [j_1 + \gamma(j_2 - j_1)]$.

$$f(\gamma) = \sum_i \log\{o_i \cdot [j_1 + \gamma(j_2 - j_1)]\}$$

$$\frac{d^2 f(\gamma)}{d\gamma^2} = -\sum_i \frac{[o_i \cdot (j_2 - j_1)]^2}{\{o_i \cdot [j_1 + \gamma(j_2 - j_1)]\}^2}$$

Since the second derivative of f is the arithmetic inverse of a sum of squares, it is never positive; hence f is concave. Since j_1 and j_2 are arbitrary JDVs, L must be concave.

Now assume that j_1 and j_2 are maximizers of L. Since the log likelihood function is concave, all $j_1 + \gamma(j_2 - j_1)$ with $0 < \gamma < 1$ also maximize the joint likelihood function. Hence, the second derivative of f is 0 for $0 < \gamma < 1$ because $f(\gamma) = \log\{L[j_1 + \gamma(j_2 - j_1)]\}$ is a constant function.

$$0 = \frac{d^2 f(\gamma)}{d\gamma^2} = -\sum_i \frac{[o_i \cdot (j_2 - j_1)]^2}{\{o_i \cdot [j_1 + \gamma(j_2 - j_1)]\}^2}$$

Clearly $\forall_i \; o_i \cdot (j_2 - j_1) = 0$; hence $j_2 - j_1$ is in the null space of our observations because it is perpendicular to all of our observation vectors. If j_1 is a maximum likelihood JDV (MLJ) then j_2 is an MLJ if and only if $j_1 - j_2$ is in the null space.

If we know that the JDV J is an MLJ for a set of experiments with observation vectors $\{o_1, o_2, \ldots, o_m\}$ and that $\{O_1, O_2, \ldots, O_{k \le m}\}$ is a minimal subset that spans the observation space, then the JDV K is an MLJ only if $\forall_{i=1}^{k} \; O_i \cdot J = O_i \cdot K$. Hence the set of MLJs can be computed by applying a set of linear constraints to the set of JDVs.

An important feature of our system is that introducing a new observable O with experiments in which O was true p of the time causes our system to assign a probability of p to O; this is in accordance with intuition. This feature derives from the fact that the $\beta(n, m)$ distribution is maximized at $(n + 1)/(m + 1)$.

7. COMPUTATIONAL METHODS

To discover a vector of positive numbers whose components sum to 1 that maximizes the likelihood function is a problem in nonlinear programming. Constraining the vector to be a probability distribution limits the search space to a convex bounded set (in the form of a hypertetrahedron). Thus we suggest applying the method of feasible directions of Topkis and Veinott [31]; this algorithm probably converges to a maximum of the polynomials I propose here.

Given a maximum likelihood joint distribution, the method of feasible directions can also discover the bounds for any proposition. If experiments have been performed testing a proposition, i.e., proposition A in our poker example, its probability is the same in all likelihood-maximizing distributions. If no experiments have been done on a proposition, such as (Not A and Not B) in the poker example, upper and lower bounds for its probability in likelihood-maximizing joint distributions can be determined by the method of feasible directions.

Given a fixed set of variables, the computational cost of evaluating the likelihood function is proportional to the number of propositions whose probabilities are bounded and on which experiments have been performed. Since adding experimental evidence makes the maxima of the likelihood function more pronounced (the absolute values of the second derivatives are never decreased), the computational cost of finding the maximum depends largely on the cost of evaluating the likelihood function. Hence entering and retrieving information is linear or better in the amount of information entered.

8. PROPOSITIONAL AXIOMS

Our methods of incorporating propositional axioms and probabilistic statements into our system are essentially identical to Nilsson's approach [26].

Statements of propositional logic, like $B \rightarrow A$, can be added to this system. They just fix the probability of certain elements of the joint distribution at 0. In particular, the value of every variable that corresponds to an element of the truth table that is false under the proposition is set to 0. Such statements reduce the size of the JDV (since probabilities that are known to be 0 need not be represented) and thus speed computation of the average MLJ.

If we added $B \rightarrow A$ to our poker example, then the probability of c would be known to be 0, and the likelihood function would be $a^{10}b(a + b)^3d^{21}(b + d)^{35}$. If we added $B \leftrightarrow A$, then the probability of b and c would be set to 0; the likelihood function would be 0 because we have experimental evidence from statement 3 that b has a positive probability. If experimental evidence directly contradicts an axiom as above, either the evidence or the axiom must be discarded.

In our system, the probability of a proposition can be limited to a specified range. This is a linear constraint on the values of the JDVs. Since the space of legitimate JDVs is the intersection of linear constraints, the method of feasible directions still applies. Thus we can insert into our system experimental evidence, axiomatic knowledge, and probability intervals.

9. CHOOSING A JDV

Section 6 characterizes the set of JDVs that maximize the likelihood of the observed evidence and characterizes the set of MLJs as a function of an arbitrary MLJ. Section 7 discusses a computationally efficient method for finding the first MLJ. Once the set of MLJs is known, techniques from Section 7 can find the maximum and minimum probabilities that an MLJ assigns to a proposition (conditional or unconditional). However, a system that returns probability intervals suffers a few drawbacks.

- Many decisions are ambiguous. Since the probability is an interval, the expected cost of a decision becomes an interval; if the expected cost intervals of two decisions overlap, it is difficult to choose between the two (minimax theories [30] may be appropriate here).
- The system is complex. Adding new information to our system changes both the probabilities in the system and the interval sizes; analyzing the result of both effects simultaneously complicates evaluation of our analysis.

It would be much simpler to find an MLJ that is best according to some metric and use the probabilities derived from that MLJ as the answer supplied by our system. If we assume that all MLJs are equally appropriate a priori, then a natural candidate for this "optimal" MLJ is the average of the set of MLJs (AMLJ), the center of gravity of the set of MLJs; hence integrating any unconditional proposition over the set of MLJs to determine its probability is equivalent to using the AMLJ.

Finding the AMLJ is the problem of finding the center of gravity of a high-dimensional convex polytope. We can define such a polytope by the set of vectors x that fit the equations $Ax = b$ and $Cx \geq d$. A encodes the fact that MLJs must differ from the one found by the method of feasible directions by an element of the null space and any equality constraints on the probability of propositions. C encodes the fact that probabilities are greater than or equal to 0 and sum to 1, and any inequality constraints (such as interval probabilities) on propositions.

Gerber [32] suggests an algorithm for finding the moments of a simplex cut off by a half-plane that, when applied for each experimental constraint, can compute the AMLJ. This algorithm is polynomial in the size of the JDV and the number of different types of experiments performed.

10. SIMPLE EXAMPLE

I will use the poker example, reprise here. Consider these two propositions:

A: Harry lit his pipe.

B: Harry has two pairs.

Some experiments about the relationship between these two events are:

1. In the first 30 hands Harry lit his pipe in 9 of them.
2. In the next 40 hands Harry had two pairs in 5 of them.
3. In the following hands you noticed that of the 6 times he lit his pipe, 5 of those times he had two pairs.

In our poker example, (0.0625, 0.2375, 0.0625, 0.6375) is the AMLJ for statements 1 and 2; we have discovered, using numerical techniques, that the MLJ for statements 1, 2, and 3 is approximately (0.174, 0.066, 0, 0.76). Since the null space for 1, 2, and 3 is itself null, this is the only MLJ. Table 1 contains the minimum, average, and maximum probabilities for these events

Table 1. Minimum, Average, and Maximum Probabilities for Events in the Poker Example

Event	1 and 2 Min : AMLJ : Max	1, 2, and 3
A	0.2 : 0.3 : 0.3	0.24
B	0.125 : 0.125 : 0.125	0.174
A and B	0 : 0.0625 : 0.125	0.174
B given A	0 : 0.208 : 0.417	0.725
B given Not A	0 : 0.089 : 0.179	0
A given B	0 : 0.5 : 1	1
A given Not B	0.2 : 0.271 : 0.343	0.080

given the JDVs above. Table 1 demonstrates several facts about our poker example.

In the case where only observations 1 and 2 have been made:

- The probabilities for statement A and statement B are appropriate given the evidence.
- Any level of correlation between A and B is possible.
- B being false substantially constrains A.
- Harry's lighting his pipe tells you very little about his hand.

In the case where observations 1, 2, and 3 have been made:

- Since A and B are connected by statement 3, their probabilities are brought closer together.
- $B \rightarrow A$.
- If Harry is lighting his pipe, definitely bet on his having two pairs.
- If Harry isn't lighting his pipe, then bet your fortune he doesn't have two pairs (this is one of the difficulties with this approach).

Our poker example demonstrates how our system balances general knowledge about the frequency of an event with inferential knowledge about an event.

To further examine this case, consider what happens if we observe another 200 hands in which Harry never has two pairs (statement 4). The JDV for this case is (0.0396257, 0.130458, 0, 0.829916). Table 2 updates Table 1 to include this assertion. Statement 4 has reduced the probability of B considerably even in the case where A is true. Harry's lighting his pipe is still a very good clue that he may have two pairs.

11. EXPERIMENTS

To validate our system we concocted a simple joint distribution and randomly sampled it, simulating typical experimental data. We then used our system to estimate probabilities and intervals given these experimental data sets

Table 2. Minimum, Average, and Maximum Probabilities for Events in the Poker Example

Event	1 and 2 Min : AMLJ : Max	1, 2, and 3	1-4
A	0.3 : 0.3 : 0.3	0.240	0.170
B	0.125 : 0.125 : 0.125	0.174	0.040
A and B	0 : 0.0625 : 0.125	0.174	0.040
B given A	0 : 0.208 : 0.417	0.725	0.233
B given Not A	0 : 0.089 : 0.179	0	0
A given B	0 : 0.5 : 1	1	1
A given Not B	0.2 : 0.271 : 0.343	0.080	0.136

and measured how accurate these probabilities and intervals were. Because we know the joint distribution (we made it up), we know the true probabilities of these events and can compare them to our system's estimated probabilities.

The joint distribution we constructed had three binary primitive variables—American (A), Red (R), and Corroded Engine (C)—as in our second example. The joint distribution we supplied is shown in Table 3. This joint distribution yields probabilities that the engine corrodes in various types of cars; we will collect random samples from this distribution and determine from our system's probability intervals and estimates that the engine corrodes for four types of cars. The results from these investigations will demonstrate the power of our system for estimating conditional probabilities from experimental data.

We performed these experiments on data randomly sampled with replacement from this joint distribution. We checked

($S1$) 20 samples for property A

($S2$) 20 samples for property R

($S3$) Five samples for property C

($S4$) The next six samples with property A for property C

($S5$) The next three samples with property R for property C

($S6$) The next two samples without property A for property C

We performed this set of experiments four times to get the results in Table 4.

Table 3. Concocted Joint Distribution

Truth Values			
American	Red	Corroded	Probability
T	T	T	0.10
T	T	F	0.05
T	F	T	0.10
T	F	F	0.25
F	T	T	0.05
F	T	F	0.10
F	F	T	0.05
F	F	F	0.30

Initially our system is not very accurate, but as our sample sizes improve (though they never grow large) the system improves in accuracy. Table 5 shows how the AMLJs grow more accurate as the evidence from $E2$, $E3$, and $E4$ is successively added to the $E1$ evidence. Table 6 show the triplet derived from our increasing body of evidence for the probability that various types of cars have corroded engines. Table 7 summarizes the accuracy of this method.

Table 7 shows that our system is not monotonic but does converge on the correct answer. When all the evidence is taken together, every interval fits about the correct answer.

12. COMPARISON WITH SIMPLIFIED KYBURG-LOUI APPROACH

Here I compare our experimental results with an alternative system that I call the *simplified Kyburg-Loui* approach. It is a simplified version of what Kyburg and Loui have proposed for handling conflicting rules of evidence [2, 4]. The essence of this approach is that to compute the conditional probability of event E given conditions $C_1, C_2, \ldots,$ we find the best

Table 4. Experimental Results

	Experiment					
Set Number	$S1$	$S2$	$S3$	$S4$	$S5$	$S6$
$E1$	7	5	1	2	2	1
$E2$	9	8	2	2	1	1
$E3$	11	7	2	2	2	0
$E4$	7	6	1	3	1	0

Table 5. Estimated AMLJs

Truth Values[a]			Probability	Experimental AMLJs			
A	*R*	*C*	1	*E*1	*E*1 + *E*2	*E*1–*E*3	*E*1–*E*4
T	T	T	0.10	0.058	0.053	0.084	0.085
T	T	F	0.05	0.042	0.081	0.074	0.081
T	F	T	0.10	0.047	0.073	0.066	0.074
T	F	F	0.25	0.203	0.194	0.226	0.185
F	T	T	0.05	0.108	0.110	0.101	0.077
F	T	F	0.10	0.042	0.081	0.074	0.081
F	F	T	0.05	0.098	0.130	0.082	0.065
F	F	F	0.30	0.402	0.278	0.293	0.351

[a] *A*, American; *R*, Red; *C*, corroded.

Table 6. Probability Intervals and Estimates

Event	True Prob.	*E*1 MIN	:AMLJ	:MAX	*E* + *E*2 MIN	:AMLJ	:MAX
C given *A* & *R*	0.667	0.411	:0.583	:1	0.246	:0.394	:1
C given *A* & Not *R*	0.286	0.162	:0.189	:0.227	0.209	:0.272	:0.391
C given Not *A* & *R*	0.333	0.565	:0.722	:1	0.403	:0.574	:1
C given Not *A* & Not *R*	0.143	0.180	:0.195	:0.213	0.265	:0.318	:0.397
Event	True Prob.	*E*1 + *E*2 + *E*3 MIN	:AMLJ	:MAX	*E*1 + *E*2 + *E*3 + *E*4 MIN	:AMLJ	:MAX
C given *A* & *R*	0.667	0.363	:0.532	:1	0.344	:0.512	:1
C given *A* & Not *R*	0.286	0.180	:0.225	:0.302	0.217	:0.285	:0.416
C given Not *A* & *R*	0.333	0.405	:0.576	:1	0.322	:0.487	:1
C given Not *A* & Not *R*	0.143	0.184	:0.219	:0.274	0.131	:0.157	:0.195

Table 7. Accuracy of Intervals and Average Probabilities

Event	*E*1 Int[a]	*E*1 Error	*E*1 + *E*2 Int	*E*1 + *E*2 Error	*E*1–*E*3 Int	*E*1–*E*3 Error	*E*1–*E*4 Int	*E*1–*E*4 Error
C given *A* & *R*	*Y*	0.084	*Y*	0.272	*Y*	0.135	*Y*	0.154
C given *A* & Not *R*	*N*	0.097	*Y*	0.013	*Y*	0.060	*Y*	0.001
C given Not *A* & *R*	*N*	−0.389	*N*	−0.241	*N*	−0.243	*Y*	−0.154
C given Not *A* & Not *R*	*N*	−0.052	*N*	−0.175	*N*	−0.077	*Y*	−0.014

[a] *Y* if the true probability lies between the maximum and minimum probabilities computed from MLJs.

Table 8. The Experiment That Yields Probabilities in the Simplified Kyburg–Loui Approach

Conditional Statement	Experimental Statement
C given A & R	C given A
C given Not A & R	C given R
C given A & Not R	C given A
C given Not A & Not R	C given Not A

compatible conditional probability that we can compute from our evidence, and we use that one.

The rules to find the best compatible conditional probability are as follows:

1. Conditional probability S_1 is compatible with S_2 if the conditions on S_1 are a subset of the conditions on S_2; A, $A \mid B$, $A \mid C$, $A \mid B \wedge C$ are compatible with $A \mid B \wedge C$.
2. Statement S_1 is better than S_2 if the conditions on S_1 are a superset of the conditions on S_2; $A \mid B \wedge C$ is better than $A \mid B$.
3. If the previous statement does not apply, then S_1 is better than S_2 if more experiments were done on S_1 than on S_2.

Kyburg and Loui choose a confidence interval from the best statement or statements, but our system avoids choosing a confidence interval by computing the conditional probability of the event from the experimental evidence. If in n trials of $\alpha \mid \theta$, m came out true, then the probability assigned to $\alpha \mid \theta$ is $(m + 1)/(n + 2)$. This probability is the mean value of the β distribution derived from assuming a uniform prior for the probability of $\alpha \mid \theta$.

Table 8 shows which of the experiments that we performed is the best for each of the conditional probabilities that we are testing. Table 9 compares the probabilities we compute from our system (ML) and those of the Kyburg–Loui (KL) system, as we accumulate evidence from our four experimental test sets.

Table 10 compares the errors from the two systems as the total evidence increases. Clearly our system converges faster to a better answer than the simplified Kyburg–Loui system.

13. INADEQUACIES AND IMPROVEMENTS

There are three major difficulties with our system (I do not further consider the simplified Kyburg–Loui approach in this paper):

1. It sometimes behaves in an unintuitive fashion.
2. When convenient, it assigns the probability of 0 to events.
3. It is not consistent with Bayesian conditionalization.

Table 9. Comparison Between Probabilities Computed Using Simplified Kyburg–Loui and Maximum Likelihood

Event	True Prob	$E1$		$E1 + E2$		$E1 + E2 + E3$		$E1$–$E4$	
		ML	KL	ML	KL	ML	KL	ML	KL
C given A & R	0.667	0.583	0.375	0.394	0.357	0.532	0.35	0.512	0.385
C given A & Not R	0.286	0.189	0.375	0.272	0.357	0.225	0.35	0.285	0.385
C given Not A & R	0.333	0.722	0.6	0.574	0.5	0.576	0.636	0.487	0.5
C given Not A & Not R	0.143	0.195	0.5	0.318	0.5	0.219	0.375	0.157	0.3

Table 10. Comparison of Errors Made by the Two Systems

Event	$E1$		$E1 + E2$		$E1 + E2 + E3$		$E1$–$E4$	
	ML	KL	ML	KL	ML	KL	ML	KL
C given A & R	0.084	0.292	0.273	0.310	0.135	0.317	0.155	0.282
C given A & Not R	0.097	−0.089	0.013	−0.071	0.060	−0.064	0.001	−0.099
C given Not A & R	−0.389	−0.267	−0.241	−0.167	−0.243	−0.303	−0.154	−0.167
C given Nt A & Not R	−0.052	−0.357	−0.174	−0.357	−0.076	−0.232	−0.014	−0.157

The initial problem will occur in any objective system because there are cases where the correct decision is an unintuitive one. Such probability paradoxes abound and can be very subtle. Thus human judgment of intuitive systems is not useful for evaluating normative probability systems. A better judgment of its effectiveness is to evaluate the effectiveness of expert systems built using this system.

The application of the maximum likelihood principle leads to the second problem. Assuming certain events are impossible often maximizes the likelihood of a set of observations (if the *impossible* event is not observed). This leads to the system making overly strong statements such as "Harry never has two pairs when his pipe is unlit." Such a strong belief implies that the system is willing to bet any sum that Harry doesn't have two pairs when his pipe is unlit; this is clearly an unwise strategy.

The maximum likelihood principle also yields the last problem. If one wants to add into our system new information, then using Bayesian updating will not generate the probabilities that adding the information directly into the system and updating its constraints or polynomial would.

The last two problems can be eliminated by discarding the maximum likelihood principle and, instead, using the likelihood function and Bayes' law to translate a prior probability distribution over JDVs into a posterior distribution of JDVs. Then the probability of any event or combination of events is computed by integrating over the posterior distribution of JDVs. This integration can be speeded by the fact that most probable JDVs in this distribution will lie near an MLJ.

I am investigating deriving the distribution over joint marginal distributions with Occam's razor: the probability of a JDV would be a function of its simplicity. Solomonoff [33] has developed methods for evaluating the simplicity of distributions and assigning probabilities based on this evaluation.

Computational cost forbids the direct application of this work to large systems with many variables because the JDVs grow exponentially with the number of random variables. Independence assumptions can break up such systems into several smaller systems that can be handled by this approach. Methods for approximate solutions of very large systems of inequalities may be applicable to our system. I expect that systems with up to 20 variables are presently computationally feasible.

14. CONCLUSION

I have proposed a computational method for propositional evidence combination given logical axioms, point probabilities, probability intervals, and experimental evidence. My system returns probability intervals that are often point probabilities; it follows a strictly Bayesian interpretation of the evidence

subject to the maximum likelihood principle. In domains where the evidence is largely objective, such as medical diagnosis or computer vision, such a system may be superior to those based on Dempster–Shafer reasoning or probability networks.

ACKNOWLEDGMENTS

I gratefully acknowledge the work of Carolyn Decusatis in editing this work and the assistance of Moises Sudit regarding nonlinear programming.

References

1. Kyburg, H. E., Higher order probabilities and intervals, Tech. Rep. TR236, Dept. of Computer Science, Univ. Rochester, November 1987.
2. Kyburg, H. E., Epistemological relevance and statistical knowledge, Tech. Rep. TR251, Dept. of Computer Science, Univ. Rochester, April 1988.
3. Kyburg, H. E., Uncertainty logics, Tech. Rep. TR337, Dept. of Computer Science, Univ. Rochester, May 1990.
4. Loui, R., Theory and computation of uncertain inference and decision, PhD Thesis, Dept. of Computer Science, Univ. Rochester, September 1987.
5. Wesley, L. P., and Hanson, A. R., The use of an evidential-based model for representing knowledge and reasoning about images in the VISIONS system, *IEEE Trans. Pattern Anal. Mach. Intell.* **4**(5); 14–25, 1982.
6. Wang, C.-H., and Srihari, S. N., A framework for object recognition in a visually complex environment and its application to locating address blocks on mail pieces, *Int. J. Comput. Vision* **2**; 119–145, 1988.
7. Grosof, B. J., An inequality paradigm for probabilistic knowledge, *Uncertainty in Artificial Intelligence*, Vol. 1 (J. F. Lemmer and L. N. Kanal, Eds.), North-Holland, New York, 1985.
8. Hummel, R. A., and Landy, M. S., A statistical viewpoint on the theory of evidence, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-10**(2); 235–247, 1988.
9. Kyburg, H. E., Bayesian and non-Bayesian evidential updating, *AI* **31**; 271–293, 1987.
10. Hau, H.-Y., and Kashyap, R. L., A unified framework for reasoning with uncertainty and its interpretation in multi-valued logics, *Proc. IEEE Syst. Man Cybern.*, 158–162, 1987.
11. Good, I. J., *Probability and the Weighing of Evidence*, Hafner, New York, 1950.

12. Good, I. J., Which comes first, probability or statistics?, *Good Thinking: The Foundations of Probability and Its Applications*, pp. 59–62, Univ. Minnesota Press, Minneapolis, 1983.

13. Pearl, J., On evidential reasoning in a hierarchy of hypotheses, *AI* **28**(1); 9–16, 1986.

14. Levitt, T. S., Bayesian inference for radar imagery based surveillance, *Uncertainty in Artificial Intelligence*, Vol. 2 (J. F. Lemmer and L. N. Kanal, Eds.), North-Holland, New York, 1986.

15. Shastri, L., Evidential reasoning in semantic networks: a formal theory and its parallel implementations, Tech. Rep. TR 166, Dept. of Computer Science, Univ. Rochester, Rochester, N.Y., September 1985.

16. Pearl, J., Markov and Bayes networks: a comparison of two graphical representations of probabilistic knowledge, Tech. Rep. R-46, Cognitive Systems Laboratory, Computer Science Dept., Univ. California, Los Angeles, September 1986.

17. Elliott, H., and Derin, H., Modeling and segmentation of noisy and textured images using Gibbs random fields, Tech. Rep., Electrical and Computer Engineering, Univ. Massachusetts at Amherst, 1984.

18. Marroquin, J. L., Probabilistic solution of inverse problems, Tech. Rep., MIT AI Laboratory, September 1985.

19. Chou, P. B., and Raman, R., On relaxation algorithms based on Markov random fields, Tech. Rep. TR 212, Dept. of Computer Science, Univ. Rochester, July 1987.

20. Cohen, F. S., and Cooper, D. B., Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields, *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-9**(2); 195–219, 1987.

21. Sher, D., Minimizing the cost of errors with a Markov random field, *Pattern Recognition Letters*, **12**(2); February 1991.

22. Lauritzen, S. L., and Spiegelhalter, D. J., Local computations with probabilities on graphical structures and their applications to expert systems, *J. Roy. Stat. Soc. B* **50**(2); 157–224, 1988.

23. Herskovits, E., and Cooper, G., Kutato: an entropy-driven system for construction of probabilistic expert systems from databases, *Proceedings of the 6th Conference on Uncertainty in AI*, 54–62, July 1990.

24. Herskovits, E., and Cooper, G., Algorithms for Bayesian belief-network precomputation, *Methods Inf. Med.* **30**; 81–89, 1991.

25. Cheeseman, P., Response to: Local computations with probabilities on graphical structures and their applications to expert systems, *J. Roy. Stat. Soc. B* **50**(2); 202, 1988.

26. Nilsson, N. J., Probabilistic logic, *AI* **28**(1), 43–70, 1986.

27. Birnbaum, A., On the foundations of statistical inference: binary experiments, *Ann. Math. Stat.* **32**; 414-435, 1961.

28. Birnbaum, A., On the foundations of statistical inference, *J. Am. Stat. Assoc.* **57**; 269-306, 1962.

29. Berger, J. O., and Wolpert, R. L., *The Likelihood Principle* (Lect. Notes—Monograph Ser., Vol. 6), Institute of Mathematical Statistics, 1984.

30. Berger, J. O., *Statistical Decision Theory*, Springer-Verlag, New York, 1985.

31. Bazaraa, M. S., and Shetty, C. M., Methods of feasible directions, *Nonlinear Programming – Theory and Algorithms*, pp. 361-434, Wiley, New York, 1979.

32. Gerber, L., The volume cut off a simplex by a half-space, *Pacific J. Math.* **94**(2); 311-313, 1981.

33. Solomonoff, R. J., A system for incremental learning based on algorithmic probability, *Proceedings of the 6th Israeli Conference on AI and Computer Vision*, 518-528, December 1989.